Procesamiento del Lenguaje Natural, Revista nº 52 marzo de 2014, pp 77-84 recibido 15-11-13 revisado 06-02-14 aceptado 07-02-14

New experiments on speaker diarization for unsupervised speaking style voice building for speech synthesis

Nuevos experimentos en diarización de locutores para creación de voces para síntesis

Beatriz Martínez-González, José Manuel Pardo, J.D. Echeverry-Correa, J. M. Montero

Grupo de Tecnología del Habla, Universidad Politécnica de Madrid
Avenida Complutense s/n 28040. Madrid
{beatrizmartinez, pardo}@die.upm.es

Resumen: El uso universal de síntesis de voz en diferentes aplicaciones requeriría un desarrollo sencillo de las nuevas voces con poca intervención manual. Teniendo en cuenta la cantidad de datos multimedia disponibles en Internet y los medios de comunicación, un objetivo interesante es el desarrollo de herramientas y métodos para construir automáticamente las voces de estilo de varios de ellos. En un trabajo anterior se esbozó una metodología para la construcción de este tipo de herramientas, y se presentaron experimentos preliminares con una base de datos multiestilo. En este artículo investigamos más a fondo esta tarea y proponemos varias mejoras basadas en la selección del número apropiado de hablantes iniciales, el uso o no de filtros de reducción de ruido, el uso de la F0 y el uso de un algoritmo de detección de música. Hemos demostrado que el mejor sistema usando un algoritmo de detección de música disminuye el error de precisión 22,36% relativo para el conjunto de desarrollo y 39,64% relativo para el montaje de ensayo en comparación con el sistema base, sin degradar el factor de mérito. La precisión media para el conjunto de prueba es 90.62% desde 76.18% para los reportajes de 99,93% para los informes meteorológicos.
Palabras clave: síntesis de voz expresiva, diarización de locutores, estilos de habla, síntesis de voz

Abstract: Universal use of speech synthesis in different applications would require an easy development of new voices with little manual intervention. Considering the amount of multimedia data available on internet and media, one interesting goal is to develop tools and methods to automatically build multi-style voices from them. In a previous paper a methodology for constructing such tools was sketched, and preliminary experiments with a multi-style database were presented. In this paper we further investigate such approach and propose several improvements to it based on the selection of the appropriate number of initial speakers, the use or not of noise reduction filters, the use of the F0 feature and the use of a music detection algorithm. We have demonstrated that the best system using music detection algorithm decreases the precision error 22.36% relative for the development set and 39.64% relative for the test set compared to the baseline, without degrading the merit factor. The average precision for the test set is 90.62% ranging from 76.18% for reportages to 99.93% for meteorology reports.
Keywords: expressive speech synthesis, speaker diarization, speaking styles, voice building

1 Introduction

Universal use of speech synthesis in different applications would require an easy development of new voices with little manual intervention. One of the goals of the Simple4all Project (Clark and King, 2012) is to create the most portable speech synthesis system possible: one that could be automatically (or with limited manual supervision) applied to many domains and tasks. In order to use speech collected from the media or from media sharing sites, speech synthesis systems must be robust to the variation of

ISSN 1135-5948

the acoustic and environmental conditions. The system must be able to robustly cope with noisy ASR–processed corpora and with challenging data such as interviews, debates, home recordings, political speeches, etc. The use of diarization techniques for speaker–turn segmentation will allow the system creating homogeneous voices from heterogeneous recordings, because the number of speakers would be automatically estimated in a fully unsupervised way, and language–independent diarization techniques automatically could provide the temporal labels of the turns of a certain speaker (Anguera et al., 2012; Pardo et al., 2012). In a previous paper (Lorenzo-Trueba et al., 2012) a methodology for constructing such tools was sketched, and preliminary experiments with a multi-style database were presented. In this paper we further investigate such approach and propose several improvements to it based on the selection of the appropriate number of initial speakers, the use or not of noise reduction filters, the use of the F0 feature and the use of a music detection algorithm. A speaker diarization system is used but, in contrast to the traditional objective of optimizing speaker segmentation and identification, our goal is to create pure clusters (speakers) that can be used to synthesize style-voices. Expressive speech synthesis is a sub-field of speech synthesis that has been drawing a lot of attention lately, as until recently there was no effort paid to increasing the adequacy of the produced voices to the task they were intended to be used in. In (Lorenzo-Trueba et al., 2013) a work to synthesize expressive voices adapting average voices to the desire style is presented. They also mention the necessity of increase the available training data for each style. In this work we aim to develop a system able to extract from different style meetings pure clusters (speakers) suitable for the voice synthesis. Therefore, we accept losing some speech segments as long as the clusters generated are purer (speech from only one speaker).

2 Database

The evaluation presented in this paper is carried out using the C–ORAL–ROM (Moreno-Sandoval et al., 2005) database. This corpus is a multi–language and multi–style database covering a wide spectrum of formal and informal speaking styles, in public and private situations.

All the languages included are Romance (French, Italian, Portuguese and Spanish), with styles ranging from formal to informal, extracted either from the media or from private spontaneous natural speaking.

In this paper, the Spanish formal media styles have been analysed: news broadcasts, sports, meteorological reports, reportages, talk–shows, scientific press and interviews. These data have been extracted from media broadcasts of different stations, and they present a great deal of variability in the recording environments and a high number of speakers (more than 200). This results in some speakers uttering only a few short sentences, making them almost irrelevant from a statistical parametrical point of view.

The number of speakers per session is variable (between 1 and 28 speakers). Table 1 summarizes average characteristics of the considered sessions for each speaking style.

The manual transcriptions of these sessions are speaker turns where we can find the speaker specified, but the segment includes also noises, silences or music (everything from the end of the previous speaker to the beginning of the next). To refine these references to include speech only segments we have force aligned the speech with the text provided also in the transcriptions using acoustic models trained from the spanish partition of TC-STAR – EPPS (European Parliament Plenary Sessions) and PARL (Spanish Parliament Plenary Sessions). Although the forced alignment helped highly to this task, it was not free from errors, and we had to correct manually some labels.

Style	# sessions	#spk/session	Time/session
Interviews	5	2-4	7-9 min
Meteorology	3	1	2-3 min
News	6	5-10	7-9 min
Reportage	6	7-28	9-12 min
Scientific press	4	3-6	8-10 min
Sports	6	1-7	7-14 min
Talk shows	11	2-8	6-11min

Table 1: Features of the speaking style sessions in the C–ORAL–ROM database.

To evaluate the implemented methods this database has been splitted into two, the development set and the test set. Both sets are composed of sessions from all the styles evaluated. Around of a third part of the database has been reserved to test

experiments. The development set is composed of 27 sessions that sum up 234.26 minutes, and the test set is composed of 14 sessions that sum up 115.17 minutes.

3 Diarization system

In previous work (Lorenzo-Trueba et al., 2012) we used a simplified version of the speaker diarization system described in Pardo et al. (2012). Instead of using three input features (MFCC, Time delay of arrival –TDOA- and F0) we only used MFCCs and we did not apply any noise filtering to the recordings. Although our usual diarization system relies also on TDOAs (Martínez-González et al., 2012), in this case, we cannot use the delay features as there is only one channel from each session.

In Figure 1 we show the modules of the system. Except the Music detection module, all of them were included in the UPM diarization system of Pardo et al. (2012).

In dotted lines, a music detection module is represented whose influence in the diarization system will be evaluated in this paper. The segments detected as music by this module are discarded from the speech segments detected by the VAD module, and, therefore, will not be assigned to any speaker.

The Wiener filter intends to reduce the background noise in the recording. Although for the Multiple Distant Microphone (MDM) task the application of this filter has proved to be positive (Wooters and Huijbregts, 2007), experiments with our database render different results which will be presented in the following sections.

The audio signal is then processed by the MFCC estimation module, where MFCC vectors of 19 components [mfcc] are calculated every 10 ms with a window of 30ms. The audio signal is also processed by the Voice Activity Detector (VAD) module which is a hybrid energy-based detector and model-based decoder. The F0 module extracts the F0 feature and adds it to the clustering module as a new stream (Pardo et al., 2012).

The following module is the segmentation and agglomerative clustering process which consists of an initialization part and an iterative segmentation and merging process. The initialization process segments the speech into K blocks (equivalent to an initial hypothesis of K speakers or clusters) uniformly distributed. Every cluster is modelled using a gaussian mixture model (GMM) initially containing a number of components that has to be specified (we use 5 for [mfcc] and 1 for [F0] streams). After the initial segmentation a set of training and re-segmenting steps is carried out using EM training and Viterbi decoding. Then the merging step takes place.

When a merging takes place the segmentation and clustering steps are repeated until a stopping criterion is reached. More information about the baseline system can be consulted in Pardo, Anguera and Wooters, (2007).

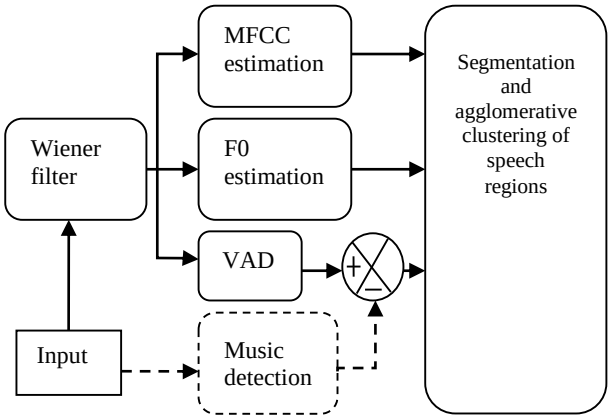

Figure 1: Block diagram of the system

4 Experiments

In this section we present new developments to the system presented in Lorenzo-Trueba et al. (2012). Different from what was presented previously is the fact that the speech/non speech transcriptions have been corrected by hand and that the database has been divided into development and test sets. The diarization score for the baseline system for the development set is included in the first row in Table 2. However, since our goal is to increase the precision of the clusters, we have calculated also the precision and recall and we have included in the last column a merit factor which weights the precision by two thirds and the recall by one third. All those values are presented in Table 2.

4.1 Initial number of speakers

The original UPM diarization system begins segmenting the recording in 16 clusters, and merging them reducing in each iteration its number. As each cluster corresponds to a hypothetical speaker, the system will never recognize more than these 16 initial speakers.

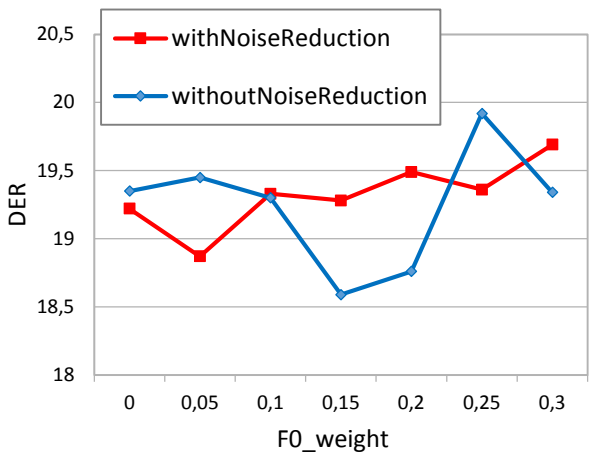

Figure 2: DER with and without applying the noise filter, using MFCC and F0 features. MFCC_weight=1-F0_weight.

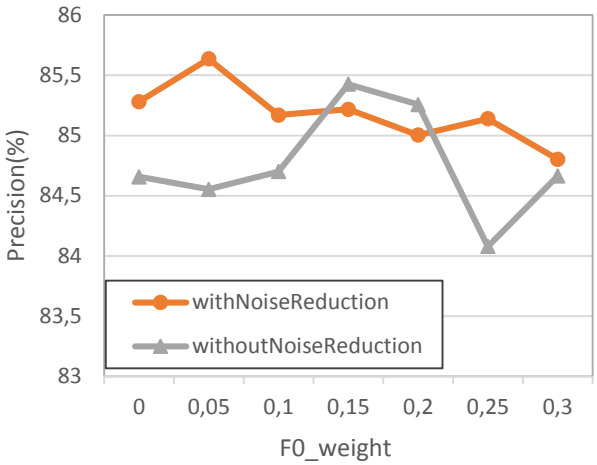

Figure 3: Precision with and without noise reduction, using MFCC and F0 features. MFCC_weight=1-F0_weight.

Some sessions have more than these 16 speakers, and thus, the system will never find all of them. In our previous experiments long sessions were splitted so no more than 9 speakers were present in a recording. In this work no sessions have been splitted so we decided to carry out some experiments beginning with 32 clusters. The best result (in precision) across different F0 weights using noise reduction (see next section) and beginning with 32 participants is shown in Table 2, second row. We noticed that even for some of the sessions with higher number of participants the results are worse than using 16 clusters (third row of Table 2). It occurs that most of the participants in the recording talked for few seconds, and these participants are hardly recognized by the system.

4.2 Noise reduction and F0

In our previous paper, we used only MFCC features to perform the diarization without noise reduction. In this work we wanted to explore the effects of applying also noise filtering and the F0 features included in Pardo et al. (2012). To combine the MFCCs with the F0 features the system needs a weight to be applied to each of these vectors. These weights are complementary, summing up 1. In Figure 2, the DER obtained for the development set when initially applying or not a noise filter (Wiener) is presented across the weight factor used for the F0 stream. The diarization error is lowest when the noise filter is not used and the weight of the F0 vector is 0.15 keeping nearly the same merit factor (see Table 2, row fourth).

Although this would be the working point in terms of DER, we had mentioned previously that our target in this diarization task is not to minimize the diarization error rate (DER) but maximize the purity of the clusters created, i.e. the precision.

The results in precision across F0 weights are shown in Figure 3. In this case the best working point is not so clear. Numerically the best precision value is obtained when the system applies noise reduction and an F0_weight of 0.05. However, this value is not so far from the best working point in the case of not applying noise reduction. In Table 2 we can see the results for both working points (rows 3rd and 4th). If we analyse the merit factor, it is very similar for both systems, so we will consider both systems in the next experiments.

4.3 Music Detection

Many of the recordings from the media have music as well as speech. The VAD module usually labels these segments as speech, and then the diarization system assigns them to one speaker, corrupting it.

If we want to use the generated clusters to synthesize voices, we want to delete any segment that would corrupt our voices. Music and noises are among the events to avoid, as well as speech overlapped with either music or noises.

System	Insertion penalty of MR module	F0 weight	DER	Precision	Recall	Merit factor	Precision error improvement (%)
Baseline		0.0	19.34	84.66	83.28	84.2	
Baseline+NR+F0+K32		0.05	20.33	84.14	82.41	83.56	-3.39
Baseline+NR+F0		0.05	18.87	85.64	83.87	85.05	6.39
Baseline+F0		0.15	18.58	85.43	84.04	84.97	5.02
Baseline+MR	**5**	**0.0**	**21.96**	**88.09**	**78.86**	**85.01**	**22.36**
Baseline+NR+F0+MR	**15**	**0.05**	**23.52**	**88.42**	**77.04**	**84.63**	**24.51**
Baseline+F0+MR	**5**	**0.15**	**21.88**	**88.17**	**78.94**	**85.09**	**22.88**
Baseline+NR+F0+MR	**5**	**0.05**	**22.49**	**88.09**	**78.24**	**84.80**	**22.35**

Table 2: Results for the development set. Relative precision error improvement is calculated over the baseline. K stands for the initial number of hypothetical speakers, K=16 if nothing indicated. NR stands for noise reduction algorithm and MR stands for music recognition algorithm.

System	Insertion penalty of MR module	F0 weight	DER	Precision	Recall	Merit factor	Precision error improvement (%)
Baseline		0.0	18.68	84.46	87.33	85.42	
Baseline+NR+F0		0.05	19.36	87.12	84.18	86.14	17.12
Baseline+F0		0.15	18.04	85.08	87.97	86.04	3.99
Baseline+MR	**5**	**0.0**	**17.28**	**90.62**	**84.38**	**88.54**	**39.64**
Baseline+NR+F0+MR	15	0.05	20.92	87.76	80.81	85.44	21.23
Baseline+F0+MR	**5**	**0.15**	**17.66**	**90.22**	**84.01**	**88.15**	**37.06**
Baseline+NR+F0+MR	5	0.05	22.57	88.95	78.92	85.60	28.89

Table 3: Results for the test set. Relative precision error improvement is calculated over the baseline. NR stands for noise reduction algorithm and MR stands for music recognition algorithm.

There are several previous works on speech and music segmentation. Many of them focus on the use of different features that would help in the discrimination between music and speech. This is the case of Izumitani, Mukai and Kashino, (2008), Gallardo-Antolin and Montero, (2010) or Panagiotakis and Tziritas, (2005). Other works like Lavner and Ruinskiy, (2009) focused in system architecture to segment speech and music.

In Gallardo and San-Segundo, (2010) the UPM-UC3M system for the Albayzin evaluation 2010 on audio segmentation is presented. The best combination of features for the segmentation of music are MFCC, CHROMA coefficients (see Bartsch and Wakefield, (2001)), and Entropy features (Misra et al., 2004). In this work we have applied this algorithm for the music segmentation.

There are five classes recognized: speech, speech+noise, speech+music, music and others. As our database is not labeled with these classes, we cannot train our own models for each of them, so, for the recognition, we used the same models that were trained in Gallardo and San-Segundo, (2010).

Once the segmentation is carried out, we only remove “music” and “others” segments from the speech segments detected by the VAD module (see diagram in Figure 1).

We carried out some experiments varying the insertion penalty in the music recognition system. The higher the term the higher the number of segments labeled as “music” or “others”.

Three kind of experiments have been carried out applying the music detection module: apply only the music detection to the baseline system, apply it in combination with F0 and in combination with F0 and the noise reduction module. For these experiments the F0 weight has been set to 0.05 when we apply noise reduction

and 0.15 when we do not (these were the two best systems in previous section, respectively rows third and fourth in Table 2).

In Figure 4, the precision and recall for the three studied systems across different insertion penalty values is presented. These three systems reach the best precision values with insertion penalty of 15 (using F0 and applying noise reduction and music detection) and 5 (for the two systems that do not use noise reduction). Higher values of this term allow more changes between classes, which means, at the end, more segments categorized as music. In fact, even if we lose more speech segments wrongly labelled as music, as long as we discard enough real music segments, the clusters generated with the remaining segments will be purer. Removing more segments, especially if they are likely to be music, could reduce the amount of speech recovered but, as long as the precision of the clusters increase and we still have enough data, the voices generated with these clusters should be more accurate. In fact, if we remove too much speech we are not only reducing the data available for voice bulding, but the models trained by the diarization system will be less accurate and, therefore, the final segmentation will have more errors.

The best numerical result (in precision) for this method is included in Table 2, sixth row (with noise reduction and insertion penalty of 15). However, in the fifth and seventh row, the best result for the two other systems with music detection are presented (no noise reduction, insertion penalty term of 5 and use or not of F0 features). We can see that even though the precision values are a bit lower, the merit factor of these two systems surpass that of the system with the best precision value (in which we applied noise reduction). The noise reduction module apparently affects highly to the recall of the system. This can be due to the high insertion penalty defined for the music detection module when using also noise reduction. For comparison purposes we have included results with the baseline, noise reduction, F0 and Music detection module when the insertion penalty is 5 (the same of the two systems without noise reduction). Precision result decreases while recall increases, but not enough to reach the performance in merit factor of any of the other two systems where no noise reduction is applied.

Our task implies maximizing precision but we want to maintain a certain level of recall and considering the variation in the merit factor we cannot yet decide between these options. Experiments with the test set will show if one of them turns clearly better.

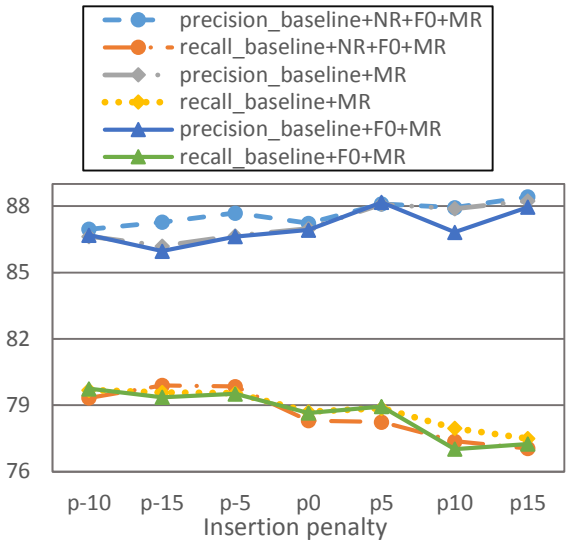

Figure 4: Precision and recall versus insertion penalty of the music recognizer for the development database. F0_weight=0.05 for system with noise reduction and 0.15 for system without it.

5 *Results with the test set and discussion*

In this section we will contrast the results of the development set with a new set, not used until now, the test set.

The first modification tried over the development set was to increase the initial number of hypothetical speakers. This modification did not improve diarization just for the development set, thus, it is not necessary a test evaluation with a different set of sessions.

The second group of experiments was focused on optimizing the systems using or not F0 and a noise reduction Wiener filter. At this point it was not clear if we should use or not the noise filtering. Both systems delivered similar performance in precision and merit factor. Thus we decide to keep both systems in future experiments.

Finally, in the last experiments with the development set, we tried to take advantage of a music detection module. This module is applied alone and in combination with the two previous ones, adjusting for each one the insertion penalty term. The three of them achieved high relative precision error improvement (24.51%, 22.88%

and 22.36%). For comparison purposes, we included also the performance of the system with the best precision result but with insertion penalty of 5 (eighth row of Table 2, 22.35% of precision error improvement).

However, these systems still had very similar precision and the best one degrades heavily its recall, and, consequently, its merit factor, so we decided to check all of them with the test set.

The experiments we have carried out with the test set to check our findings are included in Table 3.

When there is no music reduction, the use of F0 decreases the precision error in 17.12% for the system with noise reduction, which is much more than the 3.99% achieved when no noise reduction is applied (second and third row in Table 3). However, the use of noise reduction, as we have seen before, reduces heavily the recall of the system, and the merit factor of these two systems turns very similar (86.14 vs 86.04).

When we include the music detection module, the system with noise reduction (fifth and seventh row in Table 3) has the same problem we have been noticing. The recall is heavily reduced by the combination of noise reduction and music reduction, this time affecting the precision as well, which is increased much less than the two other systems with music reduction.

The two systems without noise reduction outperform clearly the rest because not only precision increases, but also the merit factor. In this case, the use of music reduction alone is slightly better than its combination with the F0 features. The precision, in this case, turns 90.62%, and recall decreases to 84.38% (vs precision 90.22% and recall of 84.01% for the system without noise reduction and F0; and precision of 87.76% and 88.95% and recall of 80.81% and 78.92% for the system with noise reduction, F0 and insertion penalty of 15 and 5 respectively), and therefore, the merit factor increases significantly.

We obtain with this system a relative decrease of the precision error of 39.64% over the test set.

We can see also, that for the test set, the use of the music reduction system decreases the DER value of the baseline in more than one point, which means that we are not discarding much clear speech, and the diarization system can model better the speakers.

Finally, in Table 4, the results obtained with different styles of the test set are presented. The precision in speaker diarization ranges from 76.18 % for reportages to 99.93% for meteorology recordings. The set of reportages is more difficult (it is the only one with precision below 90%) due to noise and the high number of different speakers that can participate (see Table 1). In future work new strategies should be drawn in order to tackle this problem.

Style	Precision	Recall	Merit factor
Interviews	92.48	91.38	92.11
Meteorology	99.93	79.18	93.01
News	96.83	93.12	95.59
Reportages	76.18	72.94	75.10
Scientific press	94.01	79.60	89.21
Sports	91.08	90.07	90.74
Talk shows	92.67	81.59	88.98
ALL	**90.62**	**84.38**	**88.54**

Table 4: Precision, recall and merit factor for the different styles in the test set.

6 Conclusions

In this paper we have analysed the task of unsupervised diarization focused on obtaining pure speaker recordings in order to synthesize voices. With this purpose we have modified slightly the traditional task of diarization. Now we have focused on recovering pure speaker clusters, even if we have to discard many segments, or speakers, overlapped with other speakers or noises. For such objective we have defined a merit factor that weights the precision and the recall. We have studied the application of some modules from the UPM diarization system and the UPM music detection module. We have proved that by using the music recognition module we can decrease the precision error 22.36% for the development set and 39.64 % for the test set, improving also the merit factor.

The noise reduction module in combination with the music reduction module makes the system to lose too many segments of speech, reducing the recall, and thus the merit factor, and making this combination undesirable.

Results using F0 in combination with music detection were slightly better for the development set and slightly worse for the test set, therefore, we cannot prove its usefulness for this task.

7 Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects.

References

Anguera, X., S. Bozonnet, N.W. D. Evans, C. Fredouille, O. Friedland, and O. Vinyals. 2012. Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*. Vol. 20, no. 2, February. ISSN: 1558-7916.

Bartsch, M. A. and G. H. Wakefield. 2001. To catch a chorus: using chroma-based representations for audio thumbnailing. *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pp.15-18.

Clark, R. and S. King. 2012, March. [Online]. Available: http://simple4all.org.

Gallardo-Antolin, A. and J. M. Montero. 2010. Histogram Equalization-Based Features for Speech, Music, and Song Discrimination. *IEEE Signal Processing Letters*, vol.17, no.7, pp.659-662, July.

Gallardo, A. and R. San-Segundo. 2010. Upm-uc3m system for music and speech segmentation. *Jornadas de Tecnología del Habla FALA 2010*. November.

Izumitani, T., R- Mukai, and K. Kashino. 2008. A background music detection method based on robust feature extraction. *IEEE International Conference on, Acoustics, Speech, and Signal Processing, 2008. ICASSP 2008*, pp. 13-16.

Lavner, Y. and D. Ruinskiy. 2009. A Decision-Tree-Based Algorithm for Speech/Music Classification and Segmentation. *EURASIP Journal on Audio, Speech, and Music Processing*.

Lorenzo-Trueba, J., B. Martínez, R. Barra-Chicote, V. López-Ludeña, J. Ferreiros, J. Yamagishi and J.M. Montero. 2012. Towards an Unsupervised Speaking Style Voice Building Framework: Multi-Style Speaker Diarization. *InterSpeech 2012*, Portland, (Oregon).

Lorenzo-Trueba, J., R. Barra-Chicote, J. Yamagishi, O. Watts and J.M. Montero. 2013. Towards Speaking Style Transplantation in Speech Synthesis. In *Proceedings SSW8 2013 - 8th ISCA Speech Synthesis Workshop*, August 31st - September 2nd.

Martínez-González, B., J. M. Pardo, J. D. Echeverry-Correa, J. A. Vallejo-Pinto and R. Barra-Chicote. 2012. Selection of TDOA parameters for MDM speaker diarization. *Interspeech 2012*, Portland (Oregon).

Misra, H., S. Ikbal, H. Bourlard and H. Hermansky. 2004. Spectral entropy based feature for robust ASR. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, vol.1, pp.I,193-6 vol.1, 17-21 May.

Moreno-Sandoval, A., G. De la Madrid, M. Alcántara, A. Gonzalez, JM Guirao and R. De la Torre, 2005. The spanish corpus, C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages. Amsterdam: John Benjamins Publishing Company, pp. 135-161.

Panagiotakis, C. and G. Tziritas. 2005. A Speech/Music Discriminator Based on RMS and Zero-Crossings. *IEEE Transactions on Multimedia, v*ol. 7, no. 1, February.

Pardo, J. M., X. Anguera, and C. Wooters. 2007. Speaker diarization for multiple-distant-microphone meetings using several sources of information. *IEEE Transactions on Computers*, vol. 56, no. 9, pp. 1212–1224, Sept.

Pardo, J. M., R. Barra-Chicote, R. San-Segundo, R. de Cordoba, and B. Martinez-Gonzalez,. 2012. Speaker diarization features: The upm contribution to the rt09 evaluation. *IEEE Transactions on Audio, Speech and Language Processing,* vol. 20, no. 2, pp. 426–435.

Wooters, C. and M. Huijbregts. 2007. The ICSI RT07s Speaker Diarization System. In *Proceedings of the Second International Workshop on Classification of Events, Activities, and Relationships (CLEAR 2007) and the Fifth Rich Transcription 2007 Meeting Recognition (RT 2007)*, Baltimore, Maryland, pp. 509-519